Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry

Reiter, L ; Claassen, M ; Schrimpf, S P ; Jovanovic, M ; Schmidt, A ; Buhmann, J M ; Hengartner, M O ; Aebersold, R

Abstract: Comprehensive characterization of a proteome is a fundamental goal in proteomics. To achieve saturation coverage of a proteome or specific subproteome via tandem mass spectrometric identification of tryptic protein sample digests, proteomics data sets are growing dramatically in size and heterogeneity. The trend toward very large integrated data sets poses so far unsolved challenges to control the uncertainty of protein identifications going beyond well established confidence measures for peptide-spectrum matches. We present MAYU, a novel strategy that reliably estimates false discovery rates for protein identifications in large scale data sets. We validated and applied MAYU using various large proteomics data sets. The data show that the size of the data set has an important and previously underestimated impact on the reliability of protein identifications. We particularly found that protein false discovery rates are significantly elevated compared with those of peptide-spectrum matches. The function provided by MAYU is critical to control the quality of proteome data repositories and thereby to enhance any study relying on these data sources. The MAYU software is available as standalone software and also integrated into the Trans-Proteomic Pipeline.

DOI: https://doi.org/10.1074/mcp.M900317-MCP200

MCP Papers in Press. Published on July 16, 2009 as Manuscript M900317-MCP200

Protein identification false discovery rates for very large proteomics datasets generated by tandem mass spectrometry

Lukas Reiter[1,2,3,4,5], Manfred Claassen[1,4,6,7], Sabine P. Schrimpf[2,5], Marko Jovanovic[2,3,5], Alexander Schmidt[4], Joachim M. Buhmann[6,7], Michael O. Hengartner[2,3,5] and Ruedi Aebersold[4,7,8,9]

1 contributed equally
2 Institute of Molecular Biology, University of Zurich, Zurich, Switzerland
3 PhD Program in Molecular Life Sciences Zurich, Zurich, Switzerland
4 Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
5 Center for Model Organism Proteomes, University of Zurich
6 Institute of Computational Science, ETH Zurich, Zurich, Switzerland
7 Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland
8 Institute for Systems Biology, Seattle WA, USA
9 Faculty of Science, University of Zurich, Zurich, Switzerland

Corresponding Author:
Prof. Ruedi Aebersold
Institute of Molecular Systems Biology
Wolfgang-Pauli-Str. 16, HPT E 78
ETH Zurich
CH-8093 Zurich
Phone: +41 44 633 31 70
Fax: +41 44 633 10 51
aebersold@imsb.biol.ethz.ch

Lukas Reiter
Institute for Molecular Systems Biology
Wolfgang-Pauli-Str. 16, HPT C 75
ETH Zurich
CH-8093 Zurich
lukas.reiter@molbio.uzh.ch

Manfred Claassen
Institute of Computational Science
Universitaetstrasse 6, CAB E 16
ETH Zurich
CH-8092 Zürich
manfred.claassen@inf.ethz.ch

Dr. Sabine P. Schrimpf
Institute of Molecular Biology
Winterthurerstrasse 190
University of Zurich
CH-8057 Zurich
sabine.schrimpf@molbio.uzh.ch

Marko Jovanovic
Institute of Molecular Biology
Winterthurerstrasse 190
University of Zurich
CH-8057 Zurich
marko.jovanovic@molbio.uzh.ch

Alexander Schmidt
Institute of Molecular Systems Biology
Wolfgang-Pauli-Str. 16
ETH Zurich
CH-8093 Zurich
schmidt@imsb.biol.ethz.ch

Prof. Joachim M. Buhmann
Institute of Computational Science
Universitätstrasse 6, CAB G 69.2
ETH Zurich
CH-8092 Zurich
jbuhmann@inf.ethz.ch

Prof. Michael O. Hengartner
Institute of Molecular Biology
Winterthurerstrasse 190
University of Zurich
CH-8057 Zurich
michael.hengartner@molbio.uzh.ch

Running Title: Protein Identification FDR

Keywords: proteomics, tandem mass spectrometry, protein identification, false discovery rate, FDR

Abbreviations: LC-MS/MS, liquid chromatography tandem mass spectrometry;
FDR, false discovery rate;
PSM, peptide-spectrum match;
PID, protein identification;
single hit, single peptide-spectrum match protein identification;
TP, true positive;
FP, false positive;
CF, protein identification containing false positive peptide-spectrum matches;
pI, isoelectric point

SUMMARY

Comprehensive characterization of a proteome is a fundamental goal in proteomics. In order to achieve saturation coverage of a proteome or specific sub proteome via tandem mass spectrometric identification of tryptic protein sample digests, proteomic data sets are growing dramatically in size and heterogeneity. The trend towards very large integrated data sets poses so far unsolved challenges to control the uncertainty of protein identifications going beyond well established confidence measures for peptide-spectrum matches. We present *MAYU*, a novel strategy that reliably estimates false discovery rates for protein identifications in large scale data sets. We validated and applied *MAYU* using various large proteomics data sets. The data show that the size of the data set has an important and previously underestimated impact on the reliability of protein identifications. We particularly find that protein false discovery rates are significantly elevated compared to those of peptide-spectrum matches. The function provided by *MAYU* is critical to control the quality of proteome data repositories and thereby to enhance any study relying on these data sources. The *MAYU* software is available as standalone software and also integrated into the trans proteomic pipeline.

INTRODUCTION

An explicit goal of proteomics is the complete description of a proteome and the measurement of its response to perturbations (Aebersold and Mann 2003). Over the last few years advances in mass spectrometry based proteomics have achieved a tremendous increase in proteome coverage (Washburn, Wolters et al. 2001; Peng, Elias et al. 2003; Omenn, States et al. 2005; Foster, de Hoog et al. 2006; King, Deutsch et al. 2006; Brunner, Ahrens et al. 2007; Baerenfaller, Grossmann et al. 2008; de Godoy, Olsen et al. 2008; Grobei, Qeli et al. 2009; Schrimpf, Weiss et al. 2009). The volume and heterogeneity of proteomic data required to substantially map out a proteome pose considerable challenges to assess the confidence of peptides and proteins that are inferred from the collected fragment ion spectra (Nesvizhskii and Aebersold 2005). While a number of statistical tools and strategies have been developed to assess the error rate of peptide-spectrum matches (PSM), estimation of the false discovery rate (FDR) of protein identifications in large datasets remains an unresolved problem. This study presents a probabilistic framework and software that addresses this issue.

The most extensive proteome coverage has generally been realized by a strategy typically referred to as shotgun proteomics. Briefly, proteins are extracted from their biological source, enzymatically digested and optionally fractionated. The resulting peptide mixtures are then analyzed by tandem mass spectrometry (MS/MS). Peptide and protein identities are inferred by computational analyses of the acquired tandem mass spectra. The data generated by shotgun proteomics experiments are highly redundant, i.e. a subset of the peptides present is repeatedly and preferentially selected for fragmentation and identified. In contrast, other subsets of peptides, e.g. those derived from low abundance proteins are more difficult to detect and a large number of fragment ion spectra have to be acquired to increase the likelihood of their detection (Brunner, Ahrens et al. 2007; Eriksson and Fenyo 2007; Mallick, Schirle et al. 2007). Consequently, proteomic studies aiming at extensive proteome coverage generate very large data sets consisting of up to millions of fragment ion spectra.

Shotgun proteomics experiments essentially aim at the compilation of a set of

reliable protein identifications covering the proteome as extensively as possible. This is achieved by firstly inferring a set of protein identifications (inference) and secondly assessing the reliability of these identifications (FDR estimation) (**Fig. 1**). Briefly, fragment ion spectra are assigned to peptide sequences by generating peptide-spectrum matches (PSMs) using one of a range of database search engines (e.g. Mascot, Sequest, X!Tandem) (Nesvizhskii, Vitek et al. 2007). Second, protein identifications are inferred from the PSMs by assembling the identified peptide sequences into proteins (Rappsilber and Mann 2002; Nesvizhskii and Aebersold 2005). Protein identifications are thus defined as assemblies of PSMs whose peptide sequences map to the same protein (**Fig. 1**). Neither PSMs nor protein identifications are perfect. Therefore it is essential to control the reliability of PSMs and protein identifications. Various approaches have been developed to estimate the reliability of PSMs (Keller, Nesvizhskii et al. 2002; Moore, Young et al. 2002; Elias and Gygi 2007; Kall, Storey et al. 2008). FDR (Benjamini and Hochberg), i.e. the expected fraction of false positive assignments, has become a widely used measure for reliability of PSMs. FDR for PSMs can be confidently estimated by means of decoy database search strategies in which the acquired fragment ion spectra are searched against a chimeric protein database containing all (target) protein sequences possibly present in the sample analyzed and an equal number of nonsense (decoy) sequences. Target-decoy strategies are particularly appealing since they constitute a generic and independent approach to validate PSMs generated by any type of identification strategy.

Protein identifications, i.e. assemblies of PSMs, are the biologically relevant outcome of a shotgun experiment. Therefore, it is highly desirable to directly control the quality of protein identifications, for example in terms of FDR. Deriving FDR for protein identifications is, however, not as obvious as determining FDR for PSMs. Because protein identifications are defined by assemblies of PSMs, errors determined at the PSM level propagate to the protein identification level in a non trivial manner. Therefore, controlling quality on the level of PSMs does not ensure quality at the (biologically relevant) level of protein identifications. This issue has so far not been appropriately appreciated, since the distinction between PSMs and protein identifications is frequently

blurred in the literature. An estimate of protein identification FDR, i.e. the expected proportion of false positive protein identifications, has to account for false positive and true positive PSMs distributing differently across the protein database. While false positive PSMs comparably distribute over all entries in the database (Elias and Gygi 2007), true positive PSMs map exclusively to the smaller subset of proteins being present in the biological sample. As a result, protein identification FDR in practise is larger than the PSM FDR (Adamski, Blackwell et al. 2005).

Number, frequency and size and heterogeneity of proteomic data sets steadily increase (Washburn, Wolters et al. 2001; Peng, Elias et al. 2003; Omenn, States et al. 2005; Foster, de Hoog et al. 2006; King, Deutsch et al. 2006; Brunner, Ahrens et al. 2007; Baerenfaller, Grossmann et al. 2008; de Godoy, Olsen et al. 2008; Schrimpf, Weiss et al. 2009). Available approaches for protein identification focus on the protein inference task and provide reasonable to good error estimates for individual experiments (typically 10-100 LC-MS/MS runs), the complexity level at which most proteomics studies operate (MacCoss, Wu et al. 2002; Nesvizhskii, Keller et al. 2003; Adamski, Blackwell et al. 2005; Weatherly, Astwood et al. 2005; Price, Lucitt et al. 2007). However, none of these approaches reliably quantifies the confidence in protein identifications in very large, integrated data sets (typically 100 or more LC-MS/MS runs), e.g. in terms of quantifying FDR for protein identifications (**Fig. 1**). To date, protein identifications in large proteomics data sets have been compiled according to heuristic criteria for which so far no quantitative confidence measures like FDR have been derived at the protein identification level (Washburn, Wolters et al. 2001; Wu, MacCoss et al. 2003; Chu, Liu et al. 2006; Foster, de Hoog et al. 2006; Brunner, Ahrens et al. 2007).

To close this gap, we developed a generic strategy enabling, for the first time, to quantify the confidence in protein identifications obtained from a wide range of inference methods (**Fig. 1**) in data sets of all sizes, especially in large to very large data sets. We refer to this approach as *MAYU* (no acronym). The approach extends the well established target-decoy strategy designed to estimate FDR at PSM level (Elias and Gygi 2007; Kall, Storey et al. 2008) to the level of protein

identifications, i.e. defined assemblies of PSMs (**Fig. 1**). We applied *MAYU* to three different data sets varying in instrumentation and species. We found that data set size has a previously underestimated impact on protein identification FDR. The strategy developed and the tool that implements it could therefore be of critical importance for the generation and quality control of large proteome datasets and data bases. The *MAYU* software and a manual are publicly available for download as standalone software and also implemented in the trans proteomic pipeline (Keller, Eng et al. 2005) (**Supplementary Note 1**).

EXPERIMENTAL PROCEDURES

Spectral data and database searching.

We analyzed three different data sets, from studies varying in MS instrumentation and underlying organism. All studies were based on multi-dimensional fractionation techniques and comprised samples from *C. elegans* (Schrimpf, Weiss et al. 2009)*, L. interrogans* and *S. pombe*. While the first data set was acquired on a low resolution LTQ instrument, the latter two were acquired on a high mass accuracy LTQ-FT instrument. The *C. elegans* project is part of the Center for Model Organism Proteomes (C-MOP) initiative (http://www.mop.unizh.ch/); the *C. elegans* proteome data are available on PeptideAtlas (http://www.peptideatlas.org/) (Desiere, Deutsch et al. 2005). We searched each data set against a composite target-decoy database using Turbo Sequest (Eng, McCormack et al. 1994) and Sequest on a Sorcerer machine (Sorcerer™-SEQUEST®, 3.10.4 release). The search results were transformed to the pepXML format and further processed using the Trans Proteomic Pipeline (Keller, Eng et al. 2005) to the level of PeptideProphet (Keller, Nesvizhskii et al. 2002) in units of experiments. The pepXML files were then further analyzed with the *MAYU* software. If a peptide existed in more than one protein sequence the hit was associated with one protein representing the gene locus (Schrimpf, Weiss et al. 2009), see also (Brunner, Ahrens et al. 2007; Baerenfaller, Grossmann et al. 2008). We performed all the database searches using a concatenated target-decoy database (Elias and Gygi 2007). As target database for

the *C. elegans* data set we chose wormpep170 (ftp://ftp.wormbase.org/pub/wormbase/). For the *L. interrogans* data set we used NC_005824 (http://www.ncbi.nlm.nih.gov/nuccore/45655585) and for the *S. pombe* data set we respecitively used 78.S_pombe (ftp://ftp.ebi.ac.uk/pub/databases/integr8/fasta/proteomes/). As decoy databases we used the reversed sequences of the target database.

Estimate of protein identification FDR.

The set of PSMs produced in the course of a proteomics experiment give rise to protein identifications. A set of PSMs mapping to the same protein sequence defines a protein identification. A protein identification is considered to be true positive, if it contains at least one true positive PSM, and false positive if all of its PSMs are false positive. This particularly implies that a protein identification that contains false positive PSMs is not necessarily false positive. In order to estimate protein identification FDR we estimate the expected number of false positive identifications within a set of protein identifications that has been assembled from a user-defined set of PSMs, e.g. from the set of PSMs at FDR=0.01.

Based on the well established assumption that false positive PSMs equally likely map to either target or decoy database, we used the number of PSMs mapping to the decoy database as an estimate for the number of false positive PSMs mapping to the target database. The PSM FDR is then estimated as the ratio of the number of PSMs pointing to decoy- and target database respectively. Considering that target and decoy database share the same protein length distribution, the expected number of protein identifications containing false positive PSMs can be estimated analogously using the number of protein identifications mapping to the decoy database (**Fig. 2b**).

We then estimate the expected number of false positive protein identifications given the inferred number of protein identifications containing false positive PSMs. If we assume that protein identifications containing false positive PSMs uniformly distribute over the target database, then the number of false positive protein identifications is hypergeometrically distributed (**Fig. 2b,** middle panel).

See also **Supplementary Method/Note 2** for details.

This relation can be seen by regarding the protein database as an urn containing balls, each representing a protein entry. Those balls that correspond to the true positive protein identifications are green while the remaining ones are white. In the urn analogy, observing *k* false positive protein identifications then corresponds to hitting *k* white balls after drawing (without replacement) as many times from the urn as we have protein identifications containing false positive PSMs.

Having specified the probability distribution of the number of false positive protein identifications as the hypergeometric distribution, the expected number of false positive protein identifications then follows as the probability weighted average (expectation value). The estimate of protein identification FDR is computed as the ratio of expected number of false positive protein identifications and the total amount of protein identifications mapping to the target database.

We also estimated single hit FDR based on the FDR estimate for the complete set of protein identifications by applying Bayes Law. Single hit FDR is thus obtained by multiplying the FDR of the complete set of protein identifications with the fraction of single hits among the decoy protein identifications divided by the fraction of single hits among the target protein identifications.

In the **Supplementary Method 2** we provide a formal statement of the underlying assumptions and a formal derivation of the individual estimates.

Simulation of non-uniformly distributed protein identifications containing false positive PSM.

We performed simulation studies to assess the robustness of *MAYU's* FDR estimates. We simulated the outcome of proteomic experiments with varying types of distributions for false positive PSM. For each simulation we first distributed a fixed number of true positive protein identifications across the protein database (comprising N entries). We distributed false positive PSM according to a truncated exponential distribution ($\sim \lambda e^{-\lambda x}$). The rate parameter

$\lambda=1/(u \cdot N)$ was chosen for different degrees of "uniformity" u. For each simulation we determined the true protein identification FDR and its *MAYU* estimate. For each seed of distributed true positive protein identifications we performed 50 simulations and report the average relative FDR deviation.

Validation of single hit FDR using isoelectric point information.

To validate our model we independently derived an FDR estimate for single hits and compared this value to the estimation of *MAYU*. We used 67 LC-MS/MS runs of experiment 15 of the *C. elegans* data set where peptides were fractionated by isoelectric focusing according to their isoelectric point (pI) (Schrimpf, Weiss et al. 2009). We used the standard deviation σ_{pI} of isoelectric point deviations pI as a quality measure for a set H of PSMs,

$$\Delta pI(i) = pI_{pr}(i) - pI_{ex}(i)$$

$$\sigma_{\Delta pI}(H) = \sqrt{\frac{1}{|H|}\sum_{i \in H}(\Delta pI(i) - m_{\Delta pI}(H))^2}$$

where $pI_{pr}(i)$ is the isoelectric point of a PSM i predicted by Bioperl (Stajich, Block et al. 2002). $pI_{ex}(i)$ corresponds to the experimentally measured isoelectric point of a PSM i, determined as the mean isoelectric point of the high confident peptides of the respective LC-MS/MS run (PSM FDR 0.01). $m_{pI}(H)$ denotes the mean of $pI_{pr}(i)$ for PSM i in H.

In order to specify the correspondence of PSM FDR and σ_{pI}, we generated a calibration curve with sets $H_{c,x}$ of PSMs of defined PSM FDR x. These sets were compiled from high confident target hits with zero FDR complemented with an appropriate amount of decoy hits to yield the designated PSM FDR. The corresponding decoy hits were sampled from a set of target-decoy PSMs featuring the designated PSM FDR. Standard deviations were computed using 20 bootstrap samples.

We estimated FDR for the set $H_{s,x}$ of single PSM protein identifications (single hits) with PSM FDR x by computing $\sigma_{pI}(H_{s,x})$ and reading out the corresponding FDR by linear interpolation of the calibration curve.

For very small PSM FDR x we observed a significant shift of $\sigma_{pI}(H_{s,x})$ compared to the calibration curve. Arguing that TP single hit peptides focus "better" (see **Fig. 4a**) in the isoelectric focusing step, we adjust $\sigma_{pI}(H_{s,x})$ to read out the FDR.

The unadjusted initial FDR estimate FDR_{ini} is used to weight the adjustment according to the initially estimated TP single hits.

$$\sigma_{\Delta pI}^{adj}(H_{s,x}) = \sigma_{\Delta pI}(H_{s,x}) + \left(\sigma_{\Delta pI}(H_{c,0}) - \sigma_{\Delta pI}(H_{s,0})\right)\cdot\left(1 - FDR_{ini}(H_{s,x})\right)$$

Validation of single hit FDR using synthetic peptides.

We generated three different sets of synthetic peptides synthesized on a microscale using the SPOT-synthesis technology (Wenschuh, Volkmer-Engert et al. 2000; Hilpert, Winkler et al. 2007). These sets were compiled as follows:

1) As positive control we randomly selected 50 peptide sequences that were identified with at least 100 PSM with a PSM FDR of zero in the search results of the complete *C. elegans* data set.

2) As negative control we randomly selected 50 peptide sequences from decoy proteins with a PSM FDR of 0.01 in the search results of the complete *C. elegans* data set.

3) As peptides of interest we randomly selected 150 peptide sequences whose PSM in the search results of the complete *C. elegans* data set were single hits.

The search results of the complete *C. elegans* data set were processed as follows. The PSM of the complete *C. elegans* data set were extracted. Ambiguous peptides, peptides longer than 18 amino acids and cysteine containing peptides were removed. *MAYU* was run on the remaining PSM and all PSM corresponding to PSM FDR of 0.01 were extracted. From these PSM the three sets were selected as described above.

For all the 250 synthetic peptides an inclusion list was generated (Schmidt, Gehlenborg et al. 2008) and measured on an LTQ-FT instrument such that the precursors corresponding to the selected PSM were targeted. The spectra were searched using SEQUEST on a Sorcerer machine (Sorcerer™-SEQUEST®, 3.10.4 release) and filtered for an FDR of 0.01 (protein identification FDR of 0.01 estimated by *MAYU*). The resulting tandem mass spectra were then normalized to total ion current and compared to the analogously processed tandem mass spectra of the *C. elegans* data set. Each peptide was attributed to a score comparing the corresponding *C. elegans* and inclusion list fragment ion

spectrum, i.e. summed difference of normalized intensities. We trained a Gaussian mixture model for TP/FP score distributions by fitting each component to the positive and respectively negative controls and then used the mixture model to estimate the expected number of FP single hits for the peptides of interest.

MAYU analysis on ProteinProphet protein identifications.

ProteinProphet was run on the pepXML files using runprophet from the trans proteomic pipeline (Keller, Eng et al. 2005) and target/decoy protein identifications of ProteinProphet were used as input for *MAYU's* protein identification FDR calculation.

RESULTS

MAYU - FDR for protein identifications.

MAYU implements a target-decoy strategy to estimate FDR for a set of protein identifications compiled from a selection of PSMs. Target-decoy strategies to estimate FDR of PSMs rely on the well established assumption that false positive PSMs uniformly distribute between target and decoy database. Consequently, PSM FDR is estimated as the ratio of PSMs mapping to the decoy and target database, respectively (**Fig. 2a**) (Elias and Gygi 2007). *MAYU* extends this approach to estimate FDR for protein identification*s*, i.e. assemblies of PSMs (**Fig. 2b**).

Prior to *MAYU* analysis, PSMs are gathered by a target-decoy database search and processed by a protein inference engine, finally yielding a set of target and decoy protein identifications (**Fig. 1**). Note that *MAYU* analysis solely aims to estimate the false discovery rate of a set of already inferred protein identifications. *MAYU* analysis is applicable to the results of any search and protein inference engine (**Fig. 5, Supplementary Fig. 2**). The following describes the *MAYU* workflow.

MAYU processes the supplied list of protein identifications to estimate their FDR. We define a false positive protein identification as being exclusively supported by false positive PSMs and no true positive PSMs. Assuming that false positive PSMs distribute uniformly over the chimeric database, the number of the decoy protein identifications provides an estimate of target protein identifications containing false positive PSMs (seven in the example shown in **Fig. 2b**). However, the actual number of false positive protein identifications (five in **Fig. 2b**) is lower than this (naïve target-decoy) estimate, as some proteins (two in **Fig. 2b)** in the target database will contain both true and false positive PSMs.

MAYU uses the number of protein identifications in the target and decoy database and the total number of protein entries in the database (11, 7 and 19 respectively in **Fig. 2b**) to estimate the expected number of false positive protein identifications in the target database (see **Methods, Supplementary Method 2 and Supplentary Note 2**).

In summary, starting from a shotgun proteomic data set searched against a target-decoy database, the *MAYU* workflow provides comprehensive and quantitative error analysis for protein identifications.

Validation of protein identification FDR estimate.

We validated the *MAYU* approach in various ways. First we assessed the robustness of the FDR estimates under violations of the underlying assumptions. Second, we validated the *MAYU* FDR estimates by comparing them with an independent approach that estimates single PSM protein identifications (single hits) FDR based on isoelectric point (pI) information from an isoelectric focusing experiment (67 LC-MS/MS runs, *C. elegans* data set). Third, we validated *MAYU's* FDR estimates by confirming single hit FDR using synthesized peptides corresponding to single hits in the complete *C. elegans* data set (1,305 LC-MS/MS runs).

We studied the robustness of our FDR estimates under deviations from the assumptions underlying the hypergeometric model. *MAYU's* protein identification FDR relies on statistics gathered from a target-decoy search, most

importantly the number of protein identifications mapping to the decoy database. Following (Elias and Gygi 2007), we assume this number to equal the number of target protein identifications containing false positive PSM. In order to estimate protein identification FDR with the hypergeometric model, we further assume that protein identifications containing false positive PSM uniformly distribute over the protein database. To closely meet this assumption *MAYU* partitions the protein database into subsets whose entries feature similar size. The protein identification FDR estimate is obtained by applying the hypergeometric model to each of these subsets (see **Methods**). The granularity of the partition does not affect the FDR estimate as long as more than ten size bins are considered (**Fig. 3a**). We further conducted simulation studies to assess how deviations from the uniformity assumption influence the *MAYU* FDR estimate. For each simulation we assumed a fixed number of true positive protein identifications and distributed false positive PSM according to a truncated geometric distribution. For each simulation we determined the true protein identification FDR and compared with the *MAYU* estimate (**Fig. 3b**). We observe that the *MAYU* estimates are not compromised, even for considerable deviations from the uniformity assumption.

We further validated the *MAYU* FDR estimates for (non-simulated) experimental data. *MAYU's* protein identification FDR estimates are ideally validated on a test data set derived from a well-defined mix of proteins. In order to capture the relevant phenomena complicating protein identification FDR estimates, a protein reference sample of defined composition covering a significant proportion of the entire protein database (e.g. 10%) would be required. Unfortunately, such a test data set is not available and would be exceedingly difficult to construct.

We therefore validated *MAYU* on a large data set providing additional information that allows us to independently derive single hit FDR gathered from an experiment of the *C. elegans* data set where peptides were separated by isoelectric point (pI) using isoelectric focusing (experiment 15, 67 LC-MS/MS runs).

We used the standard deviation of PSM pI deviations as a quality measure for a set of PSMs. This measure grows with the fraction of false positive PSM, since their pI values distribute over the complete pI range, in contrast to those of true

positive PSM clustering closely around the measured pI. By exploiting this phenomenon, we related pI information associated to PSM evidencing single hits to their quality in terms of FDR (**Methods**, **Fig. 4 a,b**). Since for single hits, PSM FDR is equivalent to the single hit FDR, we obtain a protein identification FDR estimate for the set of single hits.

MAYU analysis yielded a single hit FDR about ten fold higher than the corresponding PSM FDR of the complete set of protein identifications. We find the surprisingly high single hit FDRs obtained by *MAYU* analysis to be independently confirmed by the pI deviation method (**Fig. 4b**). We argue that the protein identification FDR estimates produced by *MAYU* are accurate in the context of typical proteomic studies in the range of 50 LC-MS/MS runs.

We also wanted to validate *MAYU's* FDR applied to the complete *C. elegans* data set, where the error propagation effects from PSM FDR to protein identification FDR are most pronounced. Since there was no pI information available for all 20 experiments we employed a different strategy. We used synthetic peptides and compared their tandem mass spectra to the tandem mass spectra from the *C. elegans* data set (see **Methods**). We generated three sets of peptides: positive controls, negative controls and peptides of interest. The analysis was performed on the complete data set filtered with a PSM FDR of 0.01.

We recorded tandem mass spectra of the synthetic peptides in a targeted way using inclusion lists and compared them to the corresponding spectra of the *C. elegans* data set. 35 peptides of the negative control (**Fig. 4c**, red), 42 peptides of the positive control (blue) and 114 peptides of our peptides of interest (grey) were identified.

We report the summed intensity differences distributions and observe that the peptides of interest show a bimodal distribution with the two apexes very close to the apexes of the positive and negative controls. Based on a Gaussian mixture model of for positive and negative controls we estimated the fraction of false positives of our peptides of interest as 0.49 which is very consistent with the estimated 0.47 of *MAYU*. Other recent studies confirm this considerable error accumulation among single hits (Grobei, Qeli et al. 2009).

We conclude that *MAYU's* estimates are accurate in the context of a very large

data set (1,305 LC-MS/MS runs). Considering the results obtained from the pI deviation method, we conclude that MAYU achieves accurate protein FDR estimates that scale well with data set size.

Comparison of protein identification FDR estimation procedures.

We compared protein identification FDR estimates of *MAYU*, ProteinProphet and the naïve target decoy approach. We studied four different subsets of the *C. elegans* data set varying in size (1, 5, 10 and 20 cumulative experiments). Protein identifications were inferred with ProteinProphet. Protein identification FDR for these identifications were then determined with *MAYU*, with the built-in functionality of ProteinProphet and the naïve target-decoy strategy.

The naïve target-decoy strategy estimates protein identification FDR analogously to PSM FDR, i.e. by approximating the expected number of false positive (FP) protein identification by the number of decoy protein identification (**Table 1**). We observe that the naïve target-decoy strategy estimate is overly pessimistic (**Fig. 5**). This is due to true positive (TP) protein identification containing FP PSMs and thus not contributing to the pool of FP protein identifications. In contrast, ProteinProphet's FDR estimates are too optimistic. For typically sized data sets of up to 50 LC-MS/MS runs ProteinProphet and naïve target-decoy still yield reasonable protein identification FDR estimates. However, the larger the data set size the more pronounced we find its discrepancy to the *MAYU* estimates. Note the difference between FDR estimate and protein inference. The foregoing comparison only aims to compare different protein identification FDR estimates, it is not suitable to assess the protein inference functionality of ProteinProphet that provides an effective prioritization of protein identifications using the principle of parsimony.

Protein identification FDR for various data sets.

Proteomic studies typically report lists of protein identifications and specify confidence in terms of FDR at PSM level. We used various data sets to study how well PSM FDR reflects the relevant confidence measure for these lists, i.e. protein identification FDR. To this end, we applied *MAYU* to several shotgun proteomics data sets, varying in MS instrumentation and studied organism (**Fig.**

6, a-c). We analyzed isoelectric focusing experiments of a *C. elegans* (Schrimpf, Weiss et al. 2009), *L. interrogans* and *S. pombe* sample. While the first data set was acquired on a low resolution LTQ instrument, the latter two were acquired on a high mass accuracy LTQ-FT instrument. Protein identifications were compiled by lexicographical protein inference including all PSM above a score threshold (see **Methods**). We observe that protein identification FDR behaves similarly for any of the data sets. Most importantly, we note that protein identification FDR is significantly elevated compared to the PSM FDR. We conclude that the PSM FDR is not generally an appropriate confidence measure for lists of protein identifications.

Accumulation of false positive protein identifications for data sets of increasing size.

Using *MAYU* we assessed the impact of data set size on protein identification FDR. For this purpose, we analyzed the currently largest shotgun proteomic data set for *C. elegans* (Schrimpf, Weiss et al. 2009) generated at the Center for Model Organism Proteomes (C-MOP). We sub sampled this data set (5,897,279 tandem mass spectra, 1,305 LC-MS/MS runs) into 20 data units of increasing size (**Fig. 6, d-f**). For each of these units we estimated the FDR of the protein identifications defined for varying PSM FDR cutoffs.

Our analysis revealed that protein identification FDR is strongly influenced by the chosen FDR of PSMs and the size of the respective data set (**Fig. 6, d,e**). For the 20 data units, protein identification FDR increases dramatically with growing PSM FDR (**Fig. 6d**). In the largest data unit, protein identification FDR is more than 20 times the corresponding PSM FDR (**Fig. 6e**).

For all data sets shown, the apparent maximal number of true positive protein identifications achievable by the respective data unit is approached already at very low PSM FDR, in the range of 0.005 (**Fig. 6, a-c,f**). This quick convergence of the expected number of TP protein identifications suggests that including less reliable PSMs mainly entails accumulation of FP protein identifications without gaining new TP protein identifications. We conclude that in order to achieve

acceptable protein identification FDR, PSMs have to be selected exceedingly stringently with increasing data set size.

DISCUSSION

MAYU is a generic strategy to estimate false discovery rates for protein identifications inferred from shotgun proteomics data sets. An implementation of *MAYU* is publicly available as standaolone software and also integrated into the trans proteomic pipeline (Keller, Eng et al. 2005) (**Supplementary Note 1**).

Unlike other well established strategies, which quantify the uncertainty of PSMs (frequently also referred to as peptide identifications), *MAYU* evaluates quality at the level of protein identifications. *MAYU* implements a novel and generic strategy that generalizes the established target-decoy database search approach for PSMs in order to estimate FDR for protein identifications. This approach constitutes a shift from assessing confidence of proteomic data sets at PSM level by providing instead a confidence measure at protein level. It should be noted that *MAYU* is not designed for protein inference, i.e. for the assembly of protein identifications. Instead *MAYU* generically assesses the reliability of protein identifications already inferred by any sequence database driven identification strategy (e.g. search engines such as Sequest, Mascot or protein inference strategies such as ProteinProphet). Besides exemplarily showing *MAYU*'s compatibility to applications such as lexicographical and ProteinProphet protein inference, we also applied *MAYU* to non-ambiguous protein inference (**Supplementary Fig. 2**). With regards to conceptual as well as computational issues, *MAYU* scales well with data set size and is particularly suited for the analysis of very large integrated data sets comprising millions of tandem mass spectra. This concept is also expected to be applicable to other high throughput experiments in biology and medicine which are characterized by indirect observations.

In this study, we assessed *MAYU* on three heterogeneous data sets including the

largest shotgun proteomics data set for *C. elegans* available to date (Schrimpf, Weiss et al. 2009). FDR estimation for protein identifications on data sets of this size has not been solved satisfactorily prior to *MAYU*. Widely used protein inference tools like ProteinProphet (Nesvizhskii, Keller et al. 2003) have proven to yield reliable error estimates on data sets at the experiment level (typically 10-50 LC-MS/MS runs) but fail to estimate accurate protein identification FDR for large data sets (**Fig. 5**). Current approaches to assemble protein identification from such large data sets rely on common sense criteria for which no quantitative confidence measure at protein identification level has been reported yet. *MAYU* overcomes this limitation by providing FDR for protein identifications in arbitrarily large data sets.

We found that data set size critically influences protein identification FDR. For the integrated data set (1,305 LC-MS/MS runs), the discrepancy in FDR rises to a more than 20-fold difference, even when stringent PSM FDR thresholds are used. Besides these results obtained for protein inference as described in the **Methods** sections, we found the same trend towards larger protein identification FDR for various other protein inference strategies.

This study aims to quantify the uncertainty of protein identifications in the context of a large-scale data set. To the best of our knowledge, this is the first study that independently confirms the scale of FDR estimates. More precisely, we showed that the scale of FDR estimates for a subset of single hit are in very good agreement with an independent method relying on experimentally acquired isoelectric points of peptides (**Fig 4a**). We also showed that *MAYU*'s protein identification FDRs are reproducible regardless of the underlying decoy database (**Supplementary Figure 1**).

Other approaches like the protein inference engine ProteinProphet have been successfully applied to estimate confidence measures for protein identifications in the context of smaller data sets. ProteinProphet relies on probability estimates of given PSMs to be false, to compute the probability of the cognate protein identification to be false. Our results show that in large data sets, certain classes of PSMs are enriched in false positive PSMs. This particularly applies to PSMs

defining single hits: Their actual proportion of false positive instances was nearly two orders of magnitude larger than the average FDR for the complete set of PSMs (data not shown). This discrepancy is not a contradiction: Because false positive PSM randomly map to a very large target-decoy database, they are prone to map to previously unoccupied protein entry and therefore give rise to a single hit. Phenomena like these complicate a reasonable estimate for false positive probabilities for single PSM and thus challenge approaches like ProteinProphet to estimate FDRs at protein level in the context of large-scale data sets (**Fig. 5**). In contrast, *MAYU* estimates protein identification FDR without relying on false positive probabilities for single hit PSM, since FDR estimates are derived solely from statistics gathered at the protein identification level.

In a similar spirit, a Poisson model has been proposed to estimate the proportion of false positive protein identifications given the number of supporting PSMs (Adamski, Blackwell et al. 2005). The parametric model requires the Poisson distribution parameter to be estimated. This estimate is obtained in a heuristic way by assuming different scenarios for the validity of single hits. This model implicitly assumes statistical independence of all PSMs. Our results indicate that this assumption does not hold in general (data not shown), which confirms the coarse approximate nature of the Poisson model.

MAYU circumvents the shortcomings of such parametric assumptions. *MAYU* exploits the underlying target-decoy database search strategy and particularly addresses the phenomenon of true positive protein identifications containing false positive PSMs. This clearly distinguishes *MAYU* from naïve target-decoy strategies that approximate the number of false positive protein identifications with the number of decoy protein identifications (Weatherly, Astwood et al. 2005). These strategies overestimate protein identification FDR since they implicitly assume that all protein identifications containing false positive PSMs are false positive (**Table 1**). In particular, the degree of protein identification FDR overestimation grows with data set size (**Fig. 5**) (Weatherly, Astwood et al. 2005).

Consider the following example where all proteins of a proteome (e.g. *E. coli*)

have been truly identified. The correct protein identification FDR would thus be zero. Due to the accumulation of false positive, i.e. decoy PSM (not invalidating the true evidence for the protein identifications) the naïve target-decoy strategy will falsely estimate an FDR differing significantly from zero. Furthermore, the naïve target-decoy estimate has the undesired property of diverging stronger the more experiments will be carried out.

MAYU's FDR builds on an estimate of the number of protein identifications containing false positive PSMs. In this study we estimate this quantity by the number of decoy protein identifications. While in principle there are other means to estimate the number of protein identifications containing false positive PSMs, *MAYU* uses target-decoy database searched data sets to estimate protein identification FDRs since this represents a well understood and well accepted strategy.

In addition, we find the assumptions underlying the target-decoy search strategy to be well met. The central assumption comprises that false positive PSMs uniformly distribute between target and decoy database. Foregoing studies have discussed and shown the general validity of the target-decoy search strategy (Elias and Gygi 2007). Recurrently occurring chemical entities (e.g. unusually modified peptides), which are not represented by the protein database, could potentially challenge the validity of target-decoy strategies since each of these give rise to false positive PSM preferably mapping to the same false peptide sequence. However, the overall balanced distribution of all false positive PSMs as well as protein identifications containing false positive PSMs is not compromised, due to the large number of such entities.

We have seen that protein length has a small and controllable effect on *MAYU*'s FDR estimates (**Fig. 3a**). We observed that deviations from the uniformity assumption regarding the distribution of protein identifications containing false positive PSM do not compromise the FDR estimates (**Fig. 3b**). We furthermore observed that *MAYU*'s FDR estimates are not dependent on the underlying type of decoy database, i.e. reversed or Markov model type (**Supplementary Fig. 1**). Most importantly, we were able to independently reproduce single hit FDR (**Fig.**

4), altogether providing a strong indication that the assumptions underlying *MAYU* analysis are reasonable and provide reliable estimates of protein identification FDR.

Throughput and sensitivity of mass spectrometers applied to proteomics are steadily increasing. Data repositories have been created to store the vast amount of mass spectrometric data (Craig, Cortens et al. 2004; Desiere, Deutsch et al. 2005; Martens, Hermjakob et al. 2005; King, Deutsch et al. 2006). These repositories constitute a cornerstone for proteomics contributing to a wide range of genome-wide studies. Well curated data repositories are a prerequisite of the success of applications like spectrum library searching (Stein 1995; Craig, Cortens et al. 2006; Lam, Deutsch et al. 2007), protein expression estimates by spectral counting (Liu, Sadygov et al. 2004) and targeted proteomics approaches based on the selection of proteotypic peptides (Kuster, Schirle et al. 2005). *MAYU* enables to more efficiently utilize existing and upcoming data sets in this context by allowing a quantitative quality control of the of protein identifications. *MAYU* is the first approach to quantify the uncertainty of protein identifications in the context of large scale data sets, thereby allowing to automatically curate proteomics repositories of steadily increasing size. We conclude that approaches like *MAYU* will significantly enhance genome-wide studies based on shotgun proteomics strategies.

REFERENCES

Adamski, M., T. Blackwell, et al. (2005). "Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project." Proteomics **5**(13): 3246-61.

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

Baerenfaller, K., J. Grossmann, et al. (2008). "Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics." Science **320**(5878): 938-41.

Benjamini, Y. and Y. Hochberg Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

Brunner, E., C. H. Ahrens, et al. (2007). "A high-quality catalog of the Drosophila melanogaster proteome." Nat Biotechnol **25**(5): 576-83.

Chu, D. S., H. Liu, et al. (2006). "Sperm chromatin proteomics identifies evolutionarily conserved fertility factors." Nature **443**(7107): 101-5.

Craig, R., J. C. Cortens, et al. (2006). "Using annotated peptide mass spectrum libraries for protein identification." J Proteome Res **5**(8): 1843-9.

Craig, R., J. P. Cortens, et al. (2004). "Open source system for analyzing, validating, and storing protein identification data." J Proteome Res **3**(6): 1234-42.

de Godoy, L. M., J. V. Olsen, et al. (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature **455**(7217): 1251-4.

Desiere, F., E. W. Deutsch, et al. (2005). "Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry." Genome Biol **6**(1): R9.

Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nat Methods **4**(3): 207-14.

Eng, J. K., A. L. McCormack, et al. (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database." J Am Soc Mass Spectrom: 976-989.

Eriksson, J. and D. Fenyo (2007). "Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs." Nat Biotechnol **25**(6): 651-5.

Foster, L. J., C. L. de Hoog, et al. (2006). "A mammalian organelle map by protein correlation profiling." Cell **125**(1): 187-99.

Grobei, M. A., E. Qeli, et al. (2009). "Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function." Genome Res.

Hilpert, K., D. F. Winkler, et al. (2007). "Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion." Nat Protoc **2**(6): 1333-49.

Kall, L., J. D. Storey, et al. (2008). "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases." J Proteome Res **7**(1): 29-34.

Keller, A., J. Eng, et al. (2005). "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." Mol Syst Biol **1**: 2005 0017.

Keller, A., A. I. Nesvizhskii, et al. (2002). "Empirical statistical model to estimate the

accuracy of peptide identifications made by MS/MS and database search." Anal Chem **74**(20): 5383-92.

King, N. L., E. W. Deutsch, et al. (2006). "Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas." Genome Biol **7**(11): R106.

Kuster, B., M. Schirle, et al. (2005). "Scoring proteomes with proteotypic peptide probes." Nat Rev Mol Cell Biol **6**(7): 577-83.

Lam, H., E. W. Deutsch, et al. (2007). "Development and validation of a spectral library searching method for peptide identification from MS/MS." Proteomics **7**(5): 655-67.

Liu, H., R. G. Sadygov, et al. (2004). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." Anal Chem **76**(14): 4193-201.

MacCoss, M. J., C. C. Wu, et al. (2002). "Probability-based validation of protein identifications using a modified SEQUEST algorithm." Anal Chem **74**(21): 5593-9.

Mallick, P., M. Schirle, et al. (2007). "Computational prediction of proteotypic peptides for quantitative proteomics." Nat Biotechnol **25**(1): 125-31.

Martens, L., H. Hermjakob, et al. (2005). "PRIDE: the proteomics identifications database." Proteomics **5**(13): 3537-45.

Moore, R. E., M. K. Young, et al. (2002). "Qscore: an algorithm for evaluating SEQUEST database search results." J Am Soc Mass Spectrom **13**(4): 378-86.

Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data: the protein inference problem." Mol Cell Proteomics **4**(10): 1419-40.

Nesvizhskii, A. I., A. Keller, et al. (2003). "A statistical model for identifying proteins by tandem mass spectrometry." Anal Chem **75**(17): 4646-58.

Nesvizhskii, A. I., O. Vitek, et al. (2007). "Analysis and validation of proteomic data generated by tandem mass spectrometry." Nat Methods **4**(10): 787-97.

Omenn, G. S., D. J. States, et al. (2005). "Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database." Proteomics **5**(13): 3226-45.

Peng, J., J. E. Elias, et al. (2003). "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome." J Proteome Res **2**(1): 43-50.

Price, T. S., M. B. Lucitt, et al. (2007). "EBP, a program for protein identification using multiple tandem mass spectrometry datasets." Mol Cell Proteomics **6**(3): 527-36.

Rappsilber, J. and M. Mann (2002). "What does it mean to identify a protein in proteomics?" Trends Biochem Sci **27**(2): 74-8.

Schmidt, A., N. Gehlenborg, et al. (2008). "An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures." Mol Cell Proteomics **7**(11): 2138-50.

Schrimpf, S. P., M. Weiss, et al. (2009). "Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes." PLoS Biol **7**(3): e48.

Stajich, J. E., D. Block, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-8.

Stein, S. E. (1995). "Chemical Substructure Identification by Mass Spectral Library Searching." J Am Soc Mass Spectrom: 644-655.

Washburn, M. P., D. Wolters, et al. (2001). "Large-scale analysis of the yeast proteome

by multidimensional protein identification technology." Nat Biotechnol **19**(3): 242-7.

Weatherly, D. B., J. A. Astwood, 3rd, et al. (2005). "A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results." Mol Cell Proteomics **4**(6): 762-72.

Wenschuh, H., R. Volkmer-Engert, et al. (2000). "Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides." Biopolymers **55**(3): 188-206.

Wu, C. C., M. J. MacCoss, et al. (2003). "A method for the comprehensive proteomic analysis of membrane proteins." Nat Biotechnol **21**(5): 532-8.

ACKNOWLEDGEMENTS

We want to thank Vinzenz Lange, Christian Müller, Lukas Müller, Thomas Fuchs and Bernd Bodenmiller for careful reading of the manuscript. Further we want to thank James Eddes, Christian Panse, the Center for Model Organism Proteomes (C-MOP) and the Functional Genomics Center Zurich (FGCZ) for support. Also many thanks to the Institute for Systems Biology in Seattle, especially Terry Farrah, Natalie Tasman and Eric Deutsch for software hosting and implementation into the TPP. This work was supported by grants from the Forschungskredit of the University of Zurich, University of Zurich Research Priority Program in Systems Biology and Functional Genomics, GEBERT-RÜF Stiftung, Swiss National Science Foundation (SNF) under Grant No. 31000-10767 and with Federal (US) funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179 and by SystemsX.ch, the Swiss initiative for systems biology. MOH is Ernst Hadorn Endowed Professor of Molecular Biology.

Figure 1. Protein inference and false discovery rate estimation. Tandem mass spectra are searched against a sequence database, where each spectrum is assigned to the best matching, i.e. highest scoring peptide sequence. These assignments are referred to as peptide-spectrum matches (PSMs). The PSM can then be filtered according to their score. The quality of the filtered PSM is usually specified in terms of PSM false discovery rates (PSM FDR). Score cutoffs for PSM are usually selected according to a user-defined maximal PSM FDR.

Alternatively the filtered PSM can firstly be assembled to protein identifications. The quality of the assignments is then assessed on the level of protein identifications. *MAYU* provides a strategy to quantify this quality in terms of protein identification FDR. Compared to PSM FDR, the protein identification FDR is a more informative quality measure since it operates on biological entities of interest, i.e. proteins.

Figure 2. *MAYU* protein identification false discovery rate estimation. Estimation of peptide-spectrum match (PSM) false discovery rate (FDR) using a target-decoy strategy **(a)** and protein identification (PID) FDR by *MAYU* **(b)**. PSM in the target database can be false positive (FP) / true positive (TP). The PSM FDR (the expected fraction of false positive target PSM) can be estimated with the number of decoy PSM being false positive by definition. The PSM FDR is currently the major measure used for quality control of mass spectrometric data sets **(a)**.

The derivation of protein identification FDR has to account for protein identifications containing false positive PSMs (CF) though not being false positive protein identifications **(b,** two proteins**)**. In order to estimate the expected number of true positive (h_{tp}) and false positive (h_{fp}) protein identifications, *MAYU* implements a hypergeometric model that takes the number of target (h_t) and decoy (h_{cf}) protein identifications and the total number of protein entries in the database (N) as input.

The hypothetical example illustrates that PSM FDR (25%) and protein identification FDR (45%) can differ largely.

Figure 3. Robustness of the false discovery rate estimates of *MAYU*. *MAYU* imposes the assumption that protein identifications containing false positive PSM uniformly distribute over the protein database. To closely meet this assumption *MAYU* operates on a partition of the protein database into subsets comprising proteins of similar size. The figure depicts how the size of the partition affects the protein identification FDR estimates for different sets of PSM defined over the complete *C. elegans* data set (**a**). Partitions with more than ten size bins yield stable FDR estimates and therefore seem to yield the desired protein size homogeneity. **(b) S**imulation studies for the complete *C. elegans* set where we explicitly distributed false positive PSM according to distributions increasingly deviating from uniformity (see **Methods**). We assessed the accuracy of the *MAYU* estimate in terms of relative deviation from the true FDR depending on the degree of uniformity of the false positive PSM distribution. The inserted plot exemplarily depicts four distributions of varying uniformity. We observe that the *MAYU* estimates do not deviate more than 1% from the true FDR (e.g. 0.2±0.002%), even for considerable deviations from the uniformity assumption.

Figure 4. Validation of the false discovery rate estimates of *MAYU*. We validated the *MAYU* false discovery rate (FDR) using two data sets of different size and with two distinct methods. We used experiment 15 (67 LC-MS/MS runs) of the *C. elegans* data set where experimental isoelectric point (pI) information of peptides were available (**a, b**) and we generated synthetic peptides to validate the FDRs of the complete *C. elegans* data set (1,305 LC-MS/MS runs) (**c**).

Using experiment 15 we derived a measure of the discrepancy between the measured and the computationally predicted pIs of peptides $\sigma_{\Delta pI}$ (see **Methods**). Sets of peptide-spectrum matches (PSMs) filtered with increasing PSM FDR up to 0.2 show an increase in $\sigma_{\Delta pI}$ (**a**, blue curve). $\sigma_{\Delta pI}$ for only the single hits is significantly higher than for all PSM over the complete range indicating that the single hit FDR is much higher compared to the PSM FDR (**a**, green and blue curve). The error bars specify standard deviations from 20 bootstraps. Using $\sigma_{\Delta pI}$

of all PSMs as a calibration curve we could estimate the single hit FDR assuming that true positive (TP) single hits are not generally different from the rest of PSMs in terms of pI (**b**). We also calculated a corrected single hit FDR (**a, b** brown curve) by making the reasonable assumption that TP single hit peptides focused better in the isoelectric focusing experiment (**a**, see offset of $\sigma_{\Delta pI}$ at zero PSM FDR between the single hits and all PSMs). We found strong consistency between the *MAYU* and independent method based on peptide pI information (**b**).

We ordered three sets of synthetic peptides corresponding to randomly picked PSMs of three different classes from the complete *C. elegans* data set (see **Methods**). We recorded tandem mass spectra of the synthetic peptides in a targeted way using inclusion lists and compared them to the corresponding spectra of the *C. elegans* data set (**c**). 35 peptides of the negative control (**c**, red), 42 peptides of the positive control (**c**, blue) and 114 peptides of our peptides of interest (**c**, grey) were identified with a stringent cutoff. We could nicely separate the distributions of positive and negative controls using the summed intensity difference (see **Methods**). Based on a Gaussian mixture model of the positive and negative controls we estimated the fraction of false positives of our peptides of interest as 0.49 which is very consistent with the estimated 0.47 of *MAYU*.

Figure 5. Comparison of different protein identification false discovery rate estimation strategies. We compared protein identification false discovery rate (FDR) estimates of *MAYU*, ProteinProphet and the naïve target-decoy strategy for four different data set sizes (1, 5, 10 and 20 experiments of the *C. elegans* data set, **a-d**). The discrepancy of the alternative FDR estimates and the *MAYU* estimates grow with data set size.

Figure 6. Protein identification false discovery rates behave similarly for data sets of different species and instruments and largely depend on the size of the data set. We applied *MAYU* to three different data sets of similar size but from different organisms and instruments (59,918 **a**, 40,008 **b**, 65,553 **c** target PSMs for a PSM FDR of 0.01). In all three data sets the protein identification false discovery rate (FDR) is roughly 5 times higher than the peptide-spectrum

match (PSM) FDR. The number of estimated true positive (TP) protein identifications reaches an apparent maximal number of identifications for very low PSM FDR (**a-c, f**).

We investigated the influence of data set size using 20 compilations from the *C. elegans* data set representing 1 to 20 cumulative experiments. The ratio of protein identification FDR to PSM FDR (protein identification FDR / PSM FDR) shows clear dependence on data set size (**d**). In the complete data set (1,305 LC-MS/MS runs) the protein identification FDR is more than 20 fold higher than the PSM FDR. For all data set sizes the protein identification FDR is elevated compared to the PSM FDR over the whole range of PSM FDR (**e**) and the apparent maximal number of TP protein identifications is reached for very stringent PSM FDR of roughly 0.005 (**f**). This data suggests that increasing the PSM FDR beyond 0.005 mainly entails an accumulation of FP protein identifications.

Table 1. Results of a target-decoy database search of the complete *C. elegans* data set

	PSMs			peptide identifications			protein identifications		
PSM FDR	target	decoy	decoy/target	target	decoy	decoy/target	target	decoy	decoy/target
0.05	954,661	47,725	0.05	117,293	36,419	0.310	16,459	14,354	0.872
0.01	795,502	7,947	0.01	82,628	6,394	0.077	11,089	4,974	0.449
0.001	614,486	614	0.001	65,779	519	0.008	8,477	506	0.060

Number of target and decoy peptide-spectrum matches, peptide identifications and protein identifications for three different PSM FDRs are shown. For peptides mapping to several protein sequences only the alphabetically first protein id was considered. For any PSM FDR, the ratio of decoy to target hits is higher for peptides and again higher for proteins. Unlike for the PSMs, this ratio is not to be mistaken for FDR for peptide or protein identifications.

Figure 1. Reiter, Claassen et al.

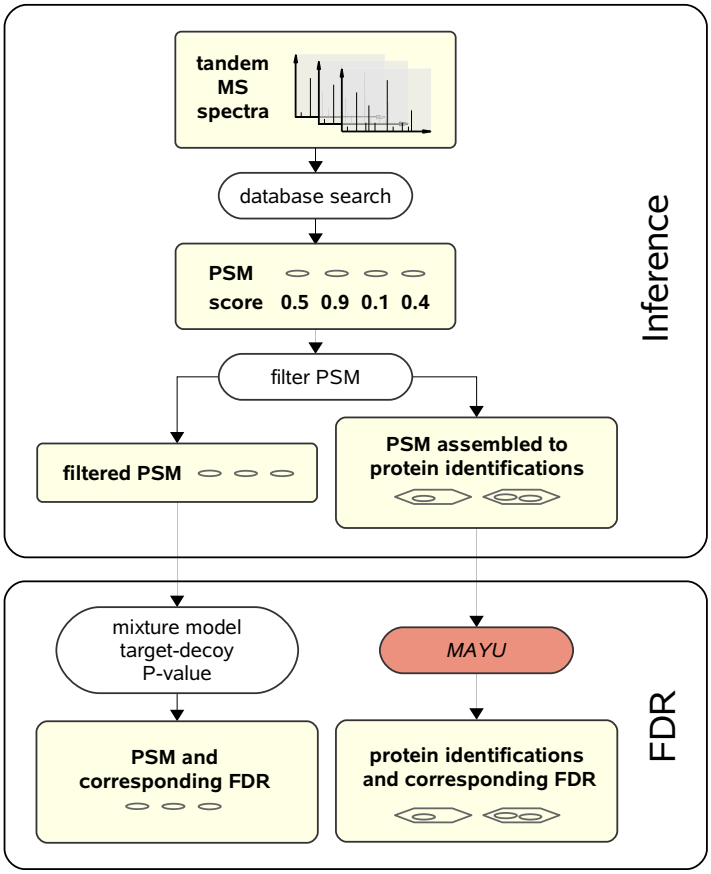

Figure 2. Reiter, Claassen et al.

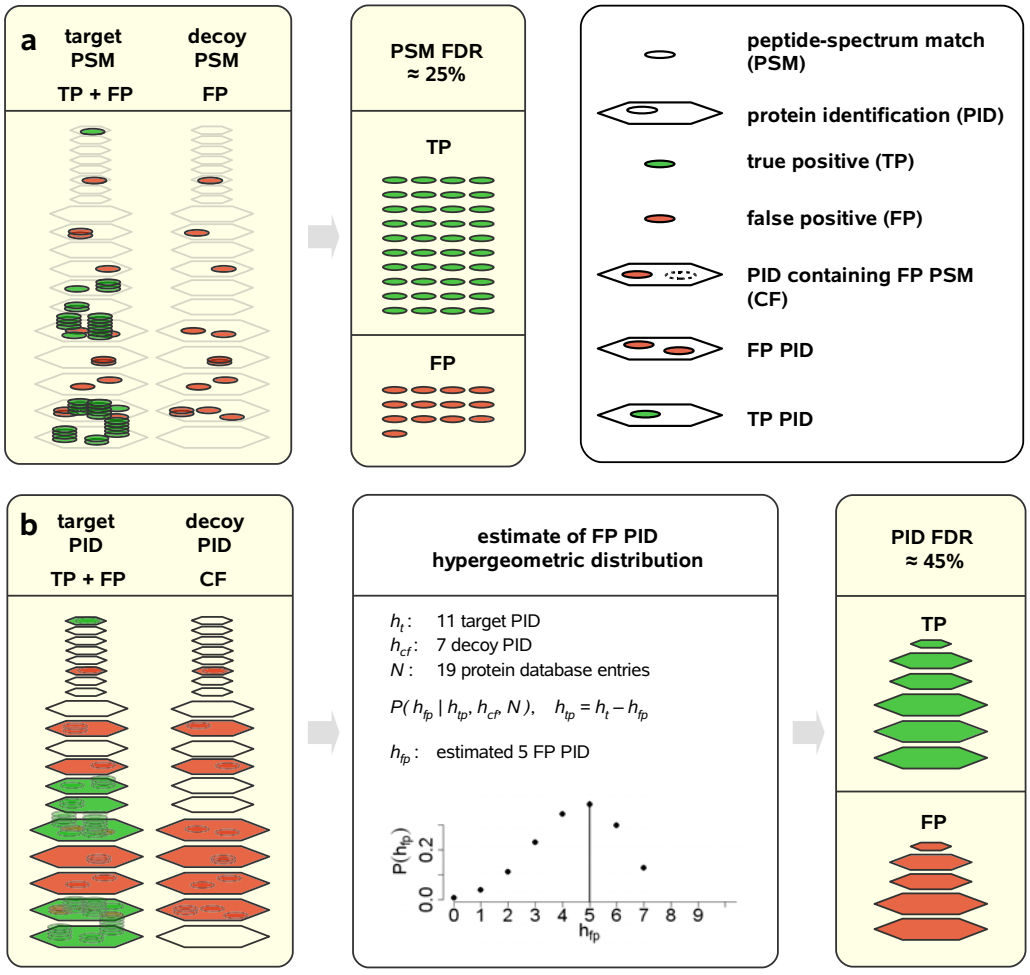

Figure 3. Reiter, Claassen et al.

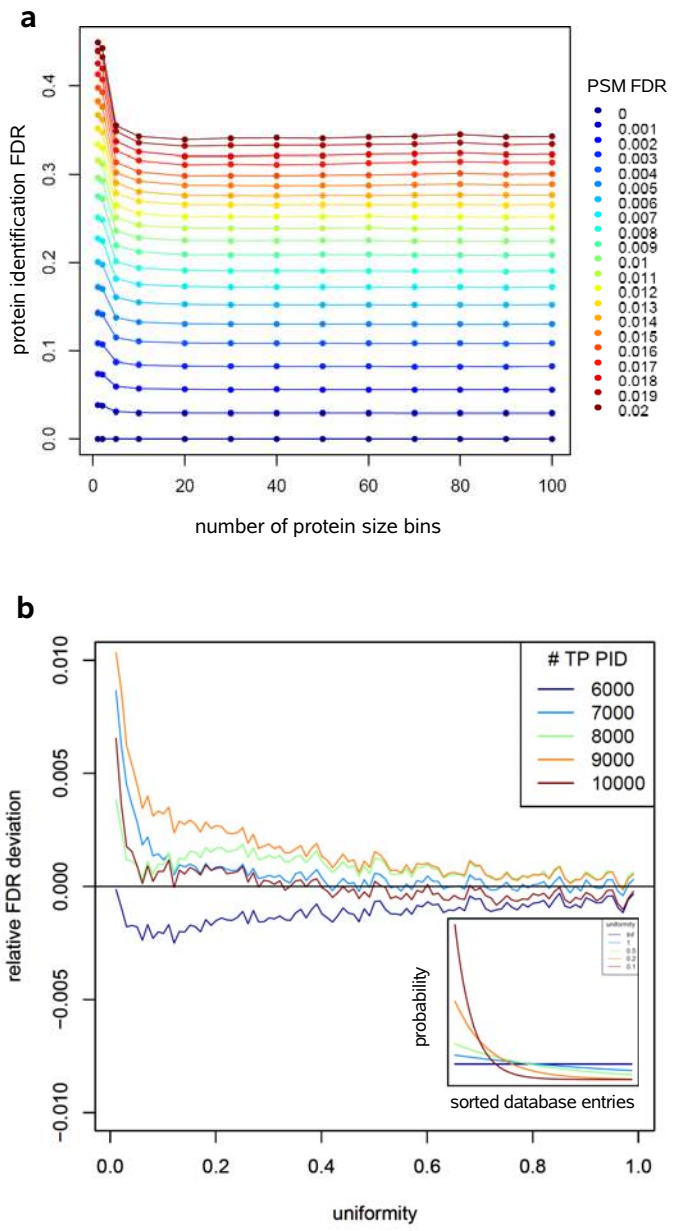

Figure 4. Reiter, Claassen et al.

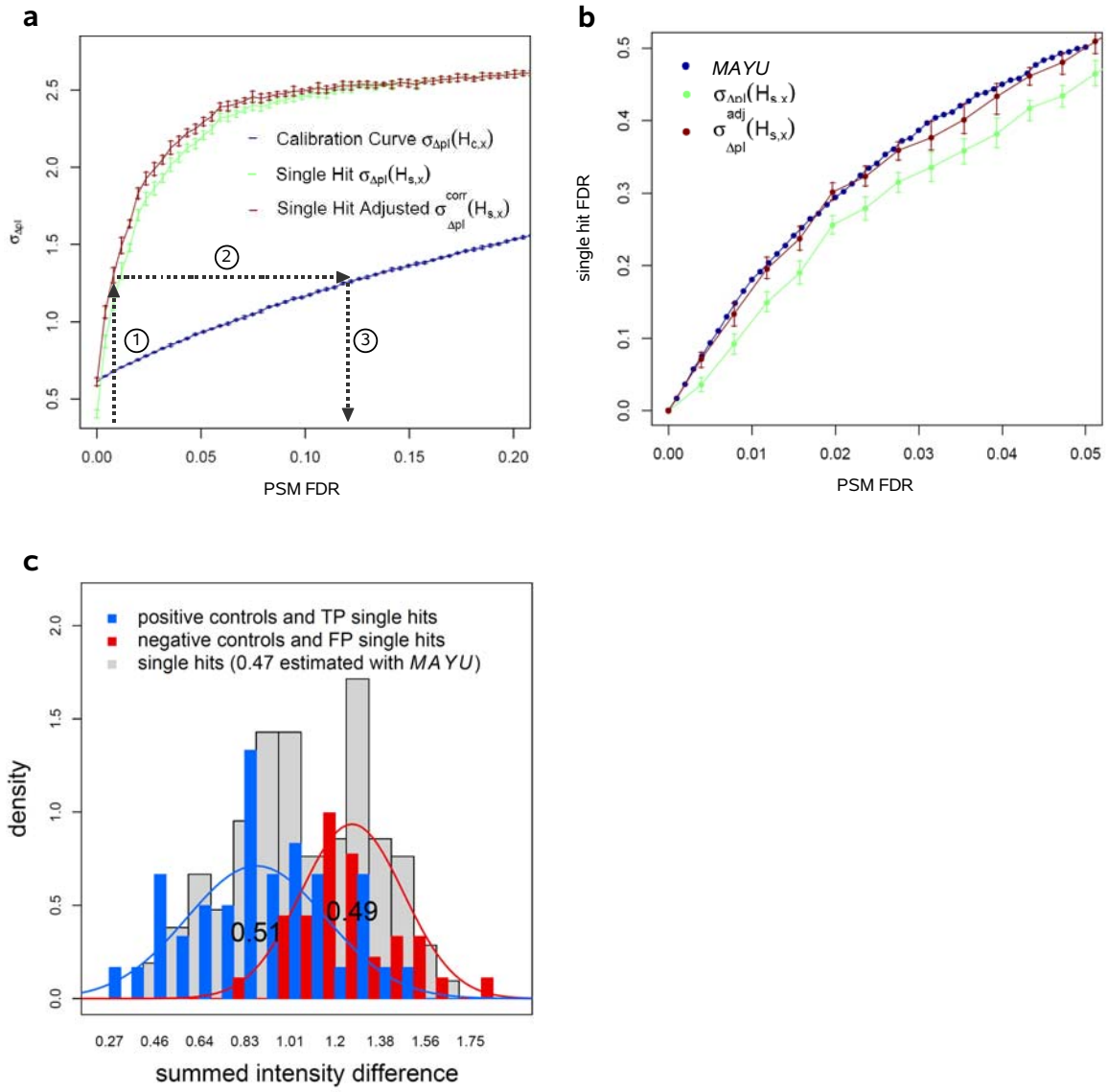

Figure 5. Reiter, Claassen et al.

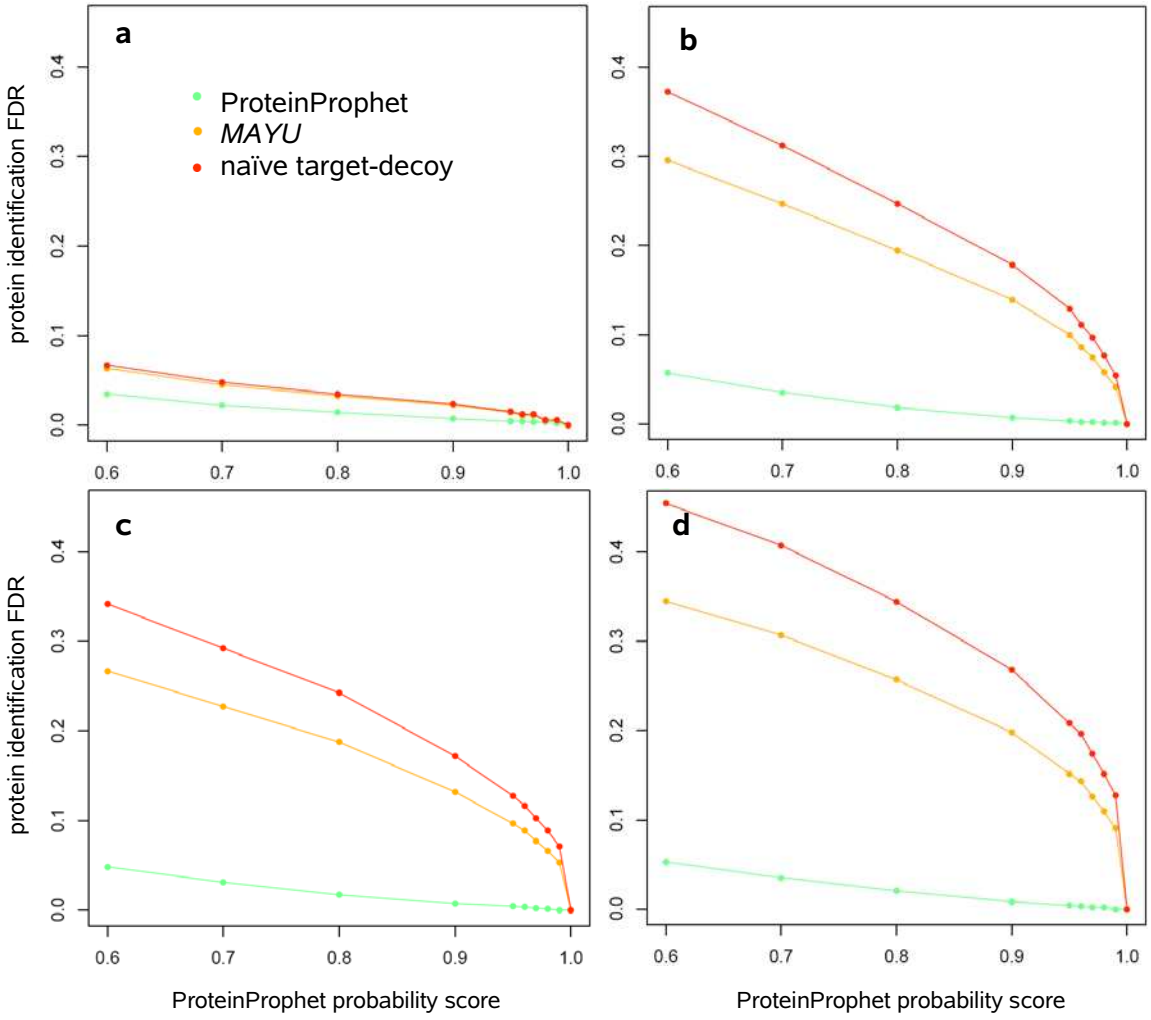

Figure 6. Reiter, Claassen et al.